A Study of Effective Information for AI-Aided Medical Diagnostics

Alexander Han

Independent Self-Paced Research

Acton-Boxborough Regional High School

Abstract

This study measures and analyzes the effectivity of typical parameters provided to the generative AI for medical diagnostics. The parameters cover symptoms, medical history, medical tests, and medications. The effectivity is measured by supplying sufficient descriptions of the parameters to AI tools GPT-3.5 and GPT-4o and analyzing the returned answers. The results show that symptoms and medical history are the most influential factors in medical diagnostics, medical tests are either vital for diagnosing certain diseases or non-factor for others, but medications are not a significant contributor.

Literature review

Chatbots have recently seen a meteoric rise, with widespread use worldwide in many fields. The most well-known chatbot is OpenAI's ChatGPT. It was launched on November 30th, 2022, and immediately caught the general public's attention. ChatGPT was not the first chatbot, but it gained the most attention due to how advanced it was compared to other chatbots. OpenAI incorporates artificial intelligence (AI) and machine learning, allowing the chatbot to respond accurately to questions [1], [2], [3], [4], [5]. At the time, GPT-3.5 contains 20 billion parameters, while GPT-3 contains 175 billion parameters but effectively performs a wide range of complex activities [6], [7], [8]. A lower amount of parameters allowed faster responses at the cost of the ability to do more complicated tasks. Recently, OpenAI introduced GPT-4 and GPT-4o, which have around 175 billion parameters and 200 billion parameters, respectively, allowing them to do much more complex functions than any other Chatbot before but at a slower rate than GPT-3.5 [9]. GPT-4o is the newest GPT version, which has significantly improved in most areas, such as accuracy, processing, and faster responses. OpenAI has allowed people to use GPT-4o without spending money for the first few questions that someone asks; it also gives an option between GPT-3.5 and GPT-4o. Since the release of ChatGPT, many other chatbots have been created that specialize in specific areas like finance, research, and medicine [10]. This study focuses on the medical chatbots. In this paper, due to the absence of formal terminology, we refer to the concept of developing an AI-based self-diagnostic tool as AI-Aided Medical Diagnostics (AAMD).

The development of AAMD can significantly benefit many individuals who are traditionally underserved by the healthcare industry. In 2021, roughly 30 million Americans (9.2% of the entire U.S. population) had no health insurance. People who do not have health insurance mostly complain about the high cost and limited coverage. In 2021, Hispanics were the minority group that would least likely have health insurance, with 30.1% of Hispanic adults having no health insurance. As a result, they would typically have to pay significantly more if they run into a medical condition. Conceivably, most of these people would recklessly avoid seeing doctors if possible. A self-diagnosis tool would enable individuals to identify potential diseases or illnesses, facilitating treatment while avoiding the high costs of professional diagnoses. Non-native English speakers can benefit from AAMD, as it is capable of understanding multiple languages. Allowing patients to self-diagnose would also provide greater privacy regarding their symptoms and encourage more active involvement in the diagnostic process, fostering a more open and collaborative relationship between patients and doctors [11].

Numerous efforts are underway to integrate AI into the medical field. In 2017, Woebot, a chatbot designed to assist with depression, anxiety, addictions, and various other aspects of mental health, was introduced to the public [12]. A distinctive feature of Woebot is its use of emojis to enhance nonverbal communication and better connect with users. Additionally, hospitals are increasingly

employing AI for medical imaging, such as CAT scans, MRIs, and X-rays, as AI can identify patterns or abnormalities that might not be easily visible to human eyes [13]. Mayo Clinic, Cleaveland Clinic, Massachusetts General Hospital, John Hopkins Hospital, and UCLA Medical Center, five of the best hospitals in the U.S., have incorporated AI into their programs. Mayo Clinic has been attempting to get AI to provide better-customized treatment options for cancer patients. In Cleaveland Clinic, the hospital is using AI to identify patients who are at high risk of cardiac arrest and need a vasopressor. Massachusetts General Hospital uses its 10 billion medical images accumulated to train the AI for radiology and pathology. John Hopkins integrates AI into the command center to improve communication between medical groups. It has also enhanced ambulance dispatch, patient triage, and patient dispatch. Finally, UCLA Medical Center deploys a chatbot, called Virtual Interventional Radiologist (VIR) that helps clinicians to reply to commonly asked questions with evidence-based answers [14]. All of these adoptions of AI could be very impactful for the future of hospitals and could significantly increase the efficiency of how a hospital works.

Despite the lingering doubts about losing jobs due to AAMD, the majority of physicians across the world are excited about implementing AI as a diagnostic tool if used properly [15]. A survey involving radiologists showed that 89% were not concerned about losing their jobs, and 77% stated that they favored adopting AI into radiology [16]. Doctors, in particular, have been excited about AI improving the diagnostic process, increasing efficiency, and improving clinical outcomes [17]. AI has already made significant strides in diagnosing different types of cancer by leveraging machine learning and its natural learning capabilities. This advancement enables doctors to provide more accurate treatments, potentially reducing cancer-related deaths substantially [10]. In the meantime, Harvard Medical School is pushing toward educating and implementing AI into healthcare as they allow students to use chatbots to help perform a diagnostic.

However, the public opinion on AAMD is almost evenly divided, with 49% favoring and 51% not favoring AI usage [18]. More specifically, individuals who oppose AAMD are primarily concerned about privacy violations and a lack of understanding regarding how AI functions. Improving patient knowledge of AI could vastly improve the public perception of AI usage in the medical field, as 65% of patients stated that they would be much more comfortable if doctors explained how AI worked in medicine and healthcare [19].

To promote the adoption of AAMD technology, ongoing efforts are required to test and document its strengths and limitations for each specific disease. An AAMD chatbot is much more likely to get a diagnosis wrong for a rare disease due to a limited training dataset. Even on GPT-4o, the bot struggles to get the correct diagnosis for rare diseases but is very reliable for common diseases like the flu or Covid-19 [20]. In addition, most of the time, the chatbots would give multiple responses rather than one clear-cut response, which hinders people from attempting to self-diagnose themselves. Another issue is knowing what knowledge the chatbot needs to diagnose correctly. Certain diseases, such as Hepatitis (a blood borne disease), can only be detected through blood tests, making self-diagnosis challenging for patients who lack the necessary materials and equipment. Another major concern is the risk of private data being exposed. As AI becomes more integrated into healthcare, a substantial amount of patient information and health records is stored online. Given the sensitive nature of health records, hackers often target them to commit fraud and remain undetected for longer periods [21].

Methodology

To reach a medical diagnosis, doctors typically ask many questions and order some lab tests if needed. Among the questions, symptoms are the central part of a diagnosis, as symptoms can significantly narrow the scope of possible diseases. Knowledge of a patient's medical history further enhances the diagnostic process, making it more precise and aiding in determining possible treatments. Lab tests are integral to the diagnostic process, providing valuable information that helps healthcare providers make informed decisions and manage patient care effectively. Lastly, the diagnostic process might be more accurate and safe when medication history is also checked.

In this study, balancing the need to include a diverse range of cases with constraints of time and expertise, three diseases—Influenza, Clostridiodes difficile (C. difficile), and Meningitis—were chosen for evaluating the AAMD tools. These diseases vary widely in their prevalence: Influenza affects approximately 1 in 16 people in the US, C. difficile occurs in about 1 in 1,000 people, and Meningitis affects around 1 in 100,000 people.

The effectivity of the four sets of parameters - symptoms, medical history, test results, and past medications- is measured by evaluating chatbots' diagnostic answers. More specifically, the effectivity consists of two scores, namely an accuracy score and a suggestiveness score. The accuracy score indicates whether the disease was among the possible conditions identified by the chatbot, while the suggestiveness score reflects how many of the suggested diagnoses are accurate.

Given the influence and popularity of various chatbots, we chose to use GPT-3.5 and GPT-4o for this research.

Table 1 shows the details of the cases used in this study. Three cases are studied for each of the three target diseases. The cases and their parameters are collected from literature and online sources. It should be noted that the "Lab Tests" column includes only the definitive tests required to confirm that AAMD tools can accurately diagnose the disease.

Diseases	Cases	Symptoms	Medical History	Medications	Lab Tests
Influenza	Case I	fever, trouble breathing	no Vaccine at all, no chronic conditions, healthy	no anti-viral treatment	Influenza PCR
	Case II	fever,trouble breathing, upset stomach, chills, muscle aches	no vaccine at all, healthy	no anti-viral treatment	Influenza PCR
	Case III	fever, increased heart rate, low blood pressure	no vaccine, mild asthma	no anti-viral treatment	Influenza PCR
C. diff	Case I	constant dirarrhea, stomach pain	visited sick grandmother with C. diff, healthy before	antibiotics for stye and parasite Blastocystis hominis	EIA stool test
	Case II	no sleep, full body muscle spasms, hot sweats, cold sweats, migraine, constant dirarrhea, stomach pain, bladder pain	get really ill if sick, sick for 8 weeks	antibiotics for sickness, 2nd antibiotics for sickness	EIA stool test
	Case III	sore throat, low body temperature, 97-102 temp	recently had colonoscopy	antibiotic for sickness	EIA stool test
Meningitis	Case I	throwing up, legs collapsed, in extreme pain	healthy before	no past medication	spinal tap
	Case II	fever, vomiting, body aches, lack of movement	healthy before	ibuprofen, Tylenol	spinal tap
	Case III	fever, headache, vision blurring, body aches, chills	healthy before	no past medication	spinal tap

Table 1 AAMD Test Cases

Results

Diagnostic parameters shown in Table 1 are transformed into questions and fed to both GPT-3.5 and GPT 4o. Based on the diagnostic answers from GPTs, the accuracy and suggestiveness scores are collected in Table 2 for GPT-3.5 and Table 3 for GPT-4o. It should be noted that the set of symptoms is chained with other sets of parameters in the chat. Moreover, the accuracy rates are 100% for the definitive lab tests of the sample diseases, which confirms that both GPT-3.5 and GPT-4o pass the basic sanity checks.

Table 2 shows that the combination of symptoms, medical history, and medicine history provides the most accurate results in GPT-3.5, with a 66.7% overall accuracy score and a 10.3% overall suggestiveness score. However, for GPT-4o (shown in Table 3), only having symptoms and medicine history gave the best results, with an overall accuracy score of 88.9% and an overall suggestiveness score of 8%. To both GPT-3.5 and GPT-4o, though, just including the symptoms alone will not render accurate enough diagnostic results, with a 44.4% accuracy score for GPT-3.5 and a 55.6% accuracy score for GPT-4o.

Diseases	Cases	Symptoms		Symptoms + Medical History		Symptoms + Medications		Symptoms + Medical History + Medications		Lab Tests
		Acc.	Sug.	Acc.	Sug.	Acc.	Sug.	Acc.	Sug.	Acc.
Influenza	Case I	100%	1 of 7	100%	1 of 6	100%	1 of 8	100%	1 of 7	100%
	Case II	100%	1 of 8	100%	1 of 6	100%	1 of 5	100%	1 of 6	100%
	Case III	0%	0 of 6	100%	1 of 7	0%	0 of 7	100%	1 of 6	100%
C. diff	Case I	0%	0 of 8	0%	0 of 7	100%	1 of 7	100%	1 of 6	100%
	Case II	0%	0 of 7	0%	0 of 8	0%	0 of 7	0%	0 of 8	100%
	Case III	0%	0 of 6	0%	0 of 4	0%	0 of 6	0%	0 of 6	100%
Meningitis	Case I	0%	0 of 5	0%	0 of 7	0%	0 of 5	0%	0 of 6	100%
	Case II	100%	1 of 8	100%	1 of 7	100%	1 of 8	100%	1 of 6	100%
	Case III	100%	1 of 8	100%	1 of 10	100%	1 of 9	100%	1 of 7	100%

Table 2 Diagnostic Accuracy and Suggestiveness Scores for GPT-3.5

Diseases	Cases	Symptoms		Symptoms + Medical History		Symptoms + Medications		Symptoms + Medical History + Medications		Lab Tests
		Acc.	Sug.	Acc.	Sug.	Acc.	Sug.	Acc.	Sug.	Acc.
Influenza	Case I	100%	1 of 12	100%	1 of 10	100%	1 of 12	100%	1 of 14	100%
	Case II	100%	1 of 12	100%	1 of 10	100%	1 of 10	100%	1 of 8	100%
	Case III	0%	0 of 12	100%	1 of 10	100%	1 of 10	100%	1 of 10	100%
C. diff	Case I	0%	0 of 16	0%	0 of 9	100%	1 of 11	100%	1 of 10	100%
	Case II	0%	0 of 10	0%	0 of 8	100%	1 of 12	0%	0 of 18	100%
	Case III	0%	0 of 12	0%	0 of 12	0%	0 of 13	0%	0 of 10	100%
Meningitis	Case I	100%	1 of 10	100%	1 of 10	100%	1 of 10	100%	1 of 12	100%
	Case II	100%	1 of 15	100%	1 of 10	100%	1 of 10	100%	1 of 10	100%
	Case III	100%	1 of 16	100%	1 of 11	100%	1 of 12	100%	1 of 10	100%

Table 3 Diagnostic Accuracy and Suggestiveness Scores for GPT-4o

The suggestiveness scores are analyzed using standard statistical methods and are illustrated in Figures 1, 2, and 3 for Influenza, C. diff, and Meningitis, respectively.

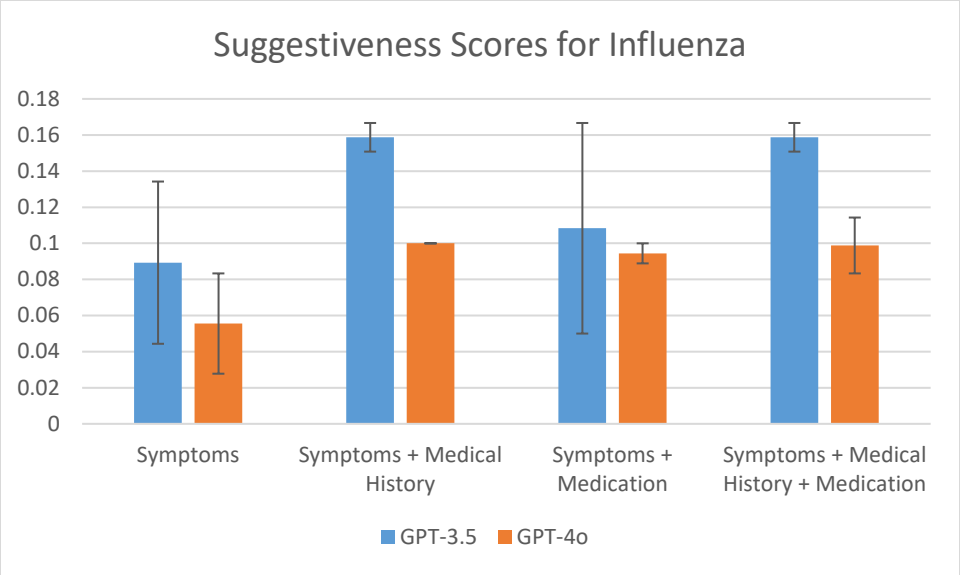

Figure 1 The Suggestiveness Scores for Influenza

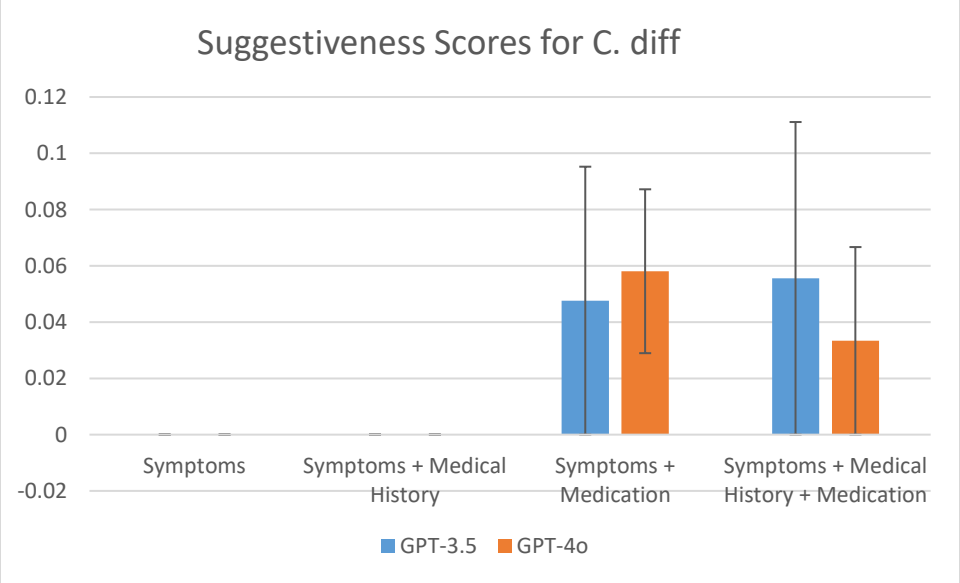

Figure 2 The Suggestiveness Scores for C. diff

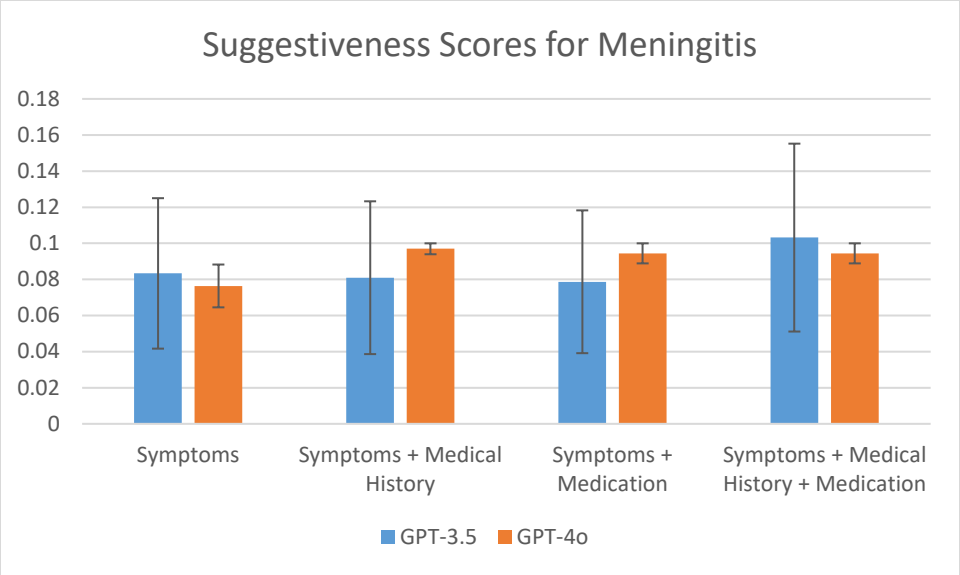

Figure 3 The Suggestiveness Scores for Meningitis

The experimental data indicates that C. Diff is challenging to diagnose for both GPT-3.5 and GPT-4o, with GPT-3.5 achieving a combined accuracy score of 16.7% and GPT-4o achieving 25%. For GPT-3.5, it diagnoses Influenza accurately, with an 83.3% accuracy score and a suggestiveness score of 12.7%. In contrast, GPT-4o shows the best results with Meningitis, attaining a perfect accuracy score of 100% and a suggestiveness score of 8.82%.

The experimental data reveals that GPT-4o has a higher chance of providing the correct diagnosis, achieving an accuracy score of 72.2%, compared to GPT-3.5's 55.6%. GPT-4o outperforms GPT-3.5 in accuracy across all three diseases tested. However, GPT-4o's suggestiveness score is lower than that of GPT-3.5, with GPT-4o scoring 6.39% versus 8.16% for GPT-3.5. On average, GPT-4o proposes 11.3 possible diagnoses, nearly doubling the 6.8 options suggested by GPT-3.5.

To determine the optimal parameter set for AAMD tools, we perform a two-way ANOVA (Analysis of Variance) test. This test evaluates how the mean suggestiveness score varies with different diagnostic parameter sets and sample diseases. Additionally, it assesses whether there is an interaction effect between these two independent variables.

Based on Tables 2 and 3, Table 4 is formed for two-way ANOVA test of the suggestiveness score with parameter sets and sample diseases as the independent variables. The ANOVA test results are shown in Table 5. We found a statistically significant difference in average suggestiveness scores by the diagnostic parameter sets ($F(2)=18.92$, $p < 0.001$), though the sample diseases and the interaction between these terms are not significant. Among the diagnostic parameter sets, the set comprising symptoms and medical history proves to be the most effective.

	Parameter Set A	Parameter Set B	Parameter Set C	Parameter Set D
Influenza	0.14	0.17	0.13	0.14
	0.13	0.17	0.20	0.17
	0.00	0.14	0.00	0.17
	0.08	0.10	0.08	0.07
	0.08	0.10	0.10	0.13
	0.00	0.10	0.10	0.10
C. diff	0.00	0.00	0.14	0.17
	0.00	0.00	0.00	0.00
	0.00	0.00	0.00	0.00
	0.00	0.00	0.09	0.10
	0.00	0.00	0.08	0.00
	0.00	0.00	0.00	0.00
Meningitis	0.00	0.00	0.00	0.00
	0.13	0.14	0.13	0.17
	0.13	0.10	0.11	0.14
	0.10	0.10	0.10	0.08
	0.07	0.10	0.10	0.10
	0.06	0.09	0.08	0.10

Table 4 Suggestiveness Scores with Parameter Sets and Sample Diseases

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	0.092019	2	0.046009	18.92106	4.25E-07	3.150411
Columns	0.015451	3	0.00515	2.117974	0.107357	2.758078
Interaction	0.013254	6	0.002209	0.908466	0.495093	2.254053
Within	0.145899	60	0.002432			
Total	0.266623	71				

Table 5 Two-way ANOVA Test Results

Discussions

The study reveals that AAMD could help patients self-diagnose themselves to some extent. GPT-3.5 and GPT-4o do well in diagnosing influenza and meningitis but struggle in diagnosing C. Diff. An area of concern is the overall suggestiveness score of GPT-4o was lower than GPT-3.5. The experiments reveal no correlation between the rarity of the disease and the effectiveness of AAMD diagnosis by either GPT-3.5 or GPT-4o. Both tools struggled to diagnose C. Diff, the middle rarity disease, while GPT-4o was perfect on the accuracy score for meningitis, the rarest disease.

It is noteworthy that excessive information may negatively impact the accuracy of the diagnosis. GPT-4o yields better scores with only symptoms and medicine history than with a full set of parameters including symptoms, medical history, and medicine history.

As previously noted, GPT-4o is more likely to include the correct diagnosis in its list but provides significantly more options than GPT-3.5. To optimize GPT-4o for accuracy while reducing the number of possible diagnoses, the question structure should be adjusted. The current question was designed for GPT-3.5, which is more open-ended, whereas GPT-4o is limited in response capacity. By removing the "List possible diseases" section from the query, the number of diagnoses generated by GPT-4o decreases. However, omitting this section from GPT-3.5 would significantly reduce its diagnostic effectiveness, as it tends to avoid vague responses and does not directly suggest possible diagnoses.

While using GPT-3.5 and GPT-4o, we observed a particularly intriguing phenomenon – GPT seems to be language dependent. If the same question is asked in different languages, the answers from GPT can be vastly different. In future research, it might be a good idea to investigate if the language is a contributing factor in AAMD.

In conclusion, this study measures and analyzes the effectivity of typical parameters provided to the generative AI for medical diagnostics. The parameters include symptoms, medical history, medical tests, and medications. Based on the accuracy scores and the suggestiveness scores, statistical analysis shows that the information of patient symptoms and medical history is the most import parameters for AI-aided medical diagnostics. The study can be extended to cover all diagnosable diseases and other AI tools. Overall, AAMD can help both patients and doctors to be more efficient in the diagnosis process, ultimately saving many lives and reducing cost.

References